We thank the reviewers for their thoughtful feedback, which will help us to improve the manuscript. We are encouraged that they found our formulation and idea to be novel(**R1**,**R3**,**R4**), intuitive (**R1**,**R2**,**R4**), impactful(**R2**), clever(**R4**), well-motivated(**R4**), and model agnostic(**R1**). We are glad they found empirical experiments to be detailed and comprehensive throughout with strong and promising results(**R1**,**R2**,**R3**,**R4**). All reviewers found the paper is well-written and results are reproducible. We are gratified that **R4** recognized the importance of AvUC loss *"differentiable approximation to the AvUC metric is pretty exciting"* and found that the *paper resoundingly addresses the problem to get reliable predictive uncertainties*. We thank the reviewers for their kind words and constructive feedback. We address reviewer comments below and will incorporate all feedback in the manuscript.

First, we would like to clarify that AvUC (accuracy vs uncertainty calibration) loss is devised to improve uncertainty calibration that can be **combined** with existing losses without modifying the underlying principles (e.g. ELBO for Bayesian NN, cross-entropy for non-Bayesian NN classifier). AvUC enables uncertainty calibration overcoming the challenge of unavailability of ground truth for uncertainty. AvUC accounts for quality of principled aleatoric and epistemic uncertainties, which are important for many applications.

R1:*Strong baseline:* Though ensemble provides higher accuracy and performs well under lower data shift intensities, we observe SVI-AvUC is more robust under high data shift intensities and out-of-distribution settings (which is common in real-world) based on the results evident from Table 1 and Fig 1,2,3. To demonstrate superiority of AvUC loss theoretically, for the camera ready version we will include a discussion section as suggested and justify how AvUC serves as loss-calibrated inference to obtain reliable uncertainties from approximate Bayesian inference (SVI) and discuss why AvUC is a proper uncertainty calibration loss in general.

Application to regression problems: Thank you for pointing us conformal prediction direction. We strongly believe our algorithm can be extended to regression problems with modifications in the formulation of loss function leveraging the relationship between prediction accuracy and uncertainty bounds. We are excited about this suggestion and we will seriously consider this for future work.

R2:*Comparison to PAvPU[35]:* Though our work is motivated from PAvPU metric, PAvPU is not a differentiable function to be used as a training loss and was originally proposed for uncertainty evaluation. We propose a trainable uncertainty calibration loss with differentiable approximation to AvU metric as defined in Eqns 4.To the best of our knowledge, this is the first work on uncertainty calibration that leverages the relationship between accuracy and uncertainty to train the model towards well-calibrated uncertainties.

R2, R3: *Theoretical justification*: Theoretically AvUC loss will be perfect 0 only when the model's uncertainty is perfectly calibrated (AvU=1). In appendix D.1, we show how AvU score and AvUC loss are related to each other during training. As noted in Eqns 3 and 4, AvUC loss attempts to maximize AvU which will indirectly push the values of uncertainties up or down based on the accuracy of predictions. When classification accuracy do not match uncertainty, $\mathrm{AvU} \rightarrow 0$ and $\mathcal{L}_{\mathrm{AvUC}} \rightarrow \infty$ forcing the gradient computation exert towards $\mathcal{L}_{\mathrm{AvUC}} \rightarrow 0$, which will happen when AvU score is pushed higher ($\mathrm{AvU} \rightarrow 1$), enabling the model to provide well-calibrated uncertainties. We will include a detailed section discussing why AvUC is a proper uncertainty calibration loss and theoretically justifying how it serves as loss-calibrated inference method (as suggested by **R4**), in the manuscript.

R3:*"Comparison with temperature scaling on top of SVI (SVI-TS):"* valid point; we have run experiments to compare with SVI-TS and will include the results for this method in manuscript. Results from the experiment for ResNet50/ImageNet in below table.

method	ECE↓ [data shift intensity (test / 1 / 2 / 3 / 4 / 5)]	UCE↓ [data shift intensity (test / 1 / 2 / 3 / 4 / 5)]
SVI-TS	0.024 / 0.030 / 0.048 / 0.072 / 0.098 / 0.123	0.117 / 0.164 / 0.199 / 0.235 / 0.269 / 0.294
SVI-AvUTS	**0.019 / 0.027 / 0.029 / 0.041 / 0.059 / 0.080**	**0.088 / 0.144 / 0.142 / 0.173 / 0.203 / 0.226**

"Comparisons have been made only on top of SVI": In appendix D.6, comparison of AvUTS on top of vanilla is presented. Also based on request from **R4**, we provide results of AvUC results applied to non-Bayesian NN here below (line 49-50). In the final version, we will include the evaluations on top of both SVI and Vanilla baselines (Bayesian and non-Bayesian method). In L194-198 of main manuscript, we mention why we chose SVI as a baseline to evaluate our methods.

Whether AvUC is a proper loss; Comparison to existing losses including MMCE[11] and cross entropy within temperature scaling: In appendix D.1, we show how AvUC loss converges based on uncertainty calibration (AvU score). When compared to existing losses including MMCE, AvUC loss accounts for quality of principled aleatoric and epistemic uncertainties while training the model, improves uncertainty calibration, can be combined with existing losses for training or used solely for post-hoc calibration and overcomes the challenge of unavailability of ground truth uncertainty. Specifically, MMCE enables confidence calibration i.e. it accounts for the probability of the predicted class, but not overall model's predictive uncertainty.

R4: Thanks for pointing the references on loss-calibrated inference methods, we agree including these citations will help to make the work more theoretically grounded and extendable as suggested. We will incorporate the feedback in final version. Regarding *training the NN solely through the AvUC loss*, we were not able to achieve training convergence in this case as AvUC is devised as an uncertainty calibration loss. However, we are able to perform temperature scaling solely with AvUC loss (AvUTS method).

Would it be possible to train a non-Bayesian NN: We trained vanilla baseline (ResNet20/CIFAR10) with AvUC loss in addition to cross-entropy as requested, the results are better (lower ECE and UCE) than solely training with cross-entropy. Also, in appendix D.6 we have presented the results of AvUTS applied to non-Bayesian NN (ResNet50/ImageNet).

method	ECE↓ [data shift intensity (test / 1 / 2 / 3 / 4 / 5)]	UCE↓ [data shift intensity (test / 1 / 2 / 3 / 4 / 5)]
Vanilla	0.046 / 0.098 / 0.139 / 0.183 / 0.236 / 0.315	0.038 / 0.085 / 0.122 / 0.162 / 0.212 / 0.285
Vanilla-AvUC	**0.034 / 0.074 / 0.102 / 0.138 / 0.185 / 0.250**	**0.017 / 0.044 / 0.064 / 0.092 / 0.134 / 0.189**

Are the results in the paper performed over different trials?: Yes, we used five different trails for CIFAR10 experiments (In Fig 2 and F9, shading in the plots shows the standard errors from different trails). For ImageNet experiments, results are reported over one trial.

Results presentation: Thank you for suggesting the presentation improvements of reported results, we will address these in camera ready version. We observed SVI-AvUC results show better spearman's ρ compared to other methods indicating least correlation of calibration errors w.r.t dataset shift [ImageNet/ECE - Vanilla:1.0, Ensemble: 0.82, SVI-AvUTS: 0.94, SVI-AvUC: **0.2**]. We will report relative ranks and include spearman rank correlation coefficient covering all the baselines, metrics and datasets in the manuscript as suggested. We have quantified the distribution separation of uncertainties in Fig 3 with Wasserstein distance [Vanilla: 2.73, Temp scaling: 2.90, Ensemble: 3.02, Dropout: 3.38, SVI: 3.73, SVI-AvUTS: 3.94, SVI-AvUC: **4.29**].